Introduction

The data supplied by the United States Census Bureau, New York State, and New York City supplied the initial data that the ETL was performed prior to drafting hypotheses.

- Completed ETL process after forming hypothesis and EDA.
- 5 questions: align with final goal which is to investigate how educational funding impacts educational outcomes by district.

Data Sources

The data sources were found by members of the team through online research. The 2019 District Data provided information about funding sources for each school district in New York (New York State, 2019). The NYC OpenData data set was used to find funding sources for NYC school districts (NYC OpenData, 2021). The NYSED website was scraped to gain extensive data about school district performance, demographics, and per-student expenditure (NYSED, 2020). Census data regarding median household income by zip code was obtained in order to find the median household income by district (United States Census Bureau, 2020). The NCES School District Geographic Relationship Files (NCES, 2021) are used to map zip codes to school districts. All data sets were used in the ETL process.

Extraction

Where did you get the data from? How did you get the data? What format is the extracted data? What steps were taken to extract the data? Be sure to number steps when the order matters.

Initial extraction

The NYSED website needed to be scraped in order to obtain the data needed.

Process for extracting data with web scraping:

1. In Jupyter Notebooks, import requests, time, BeautifulSoup, and Pandas.
2. For each district collect the raw district name in an html format, skipping empty names. When a new district name is found append it.
3. For each district collect the raw graduation and dropout rates using BeautifulSoup. Clean graduation rate and dropout rate by finding the values with the data label “Percentage of Graduates” and “Percentage of Dropout” respectively. Append these values to the graduationratelist and the dropoutratelist.
4. For each district collect the funding per student using BeautifulSoup by finding the items with the “bullet-item columns” class. Assign the value “N/A” to empty values.
5. For each district collect the number of students enrolled per district using BeautifulSoup by finding the items with the “highlight blue spacing-top-15” class.

6. For each district collect the amount of students that were male per district using BeautifulSoup by finding the items with the “small-12 large-6 medium-6 bullet-item columns” class and appending the value with an index of 0.
7. For each district collect the amount of students that were female per district using BeautifulSoup by finding the items with the “small-12 large-6 medium-6 bullet-item columns” class and appending the value with an index of 2.
8. For each district collect the number of students that were white per district using BeautifulSoup by finding the items with the "WHITE” text and appending the value. If value is empty, append “N/A”.
9. Put all of the variables into a data frame.
10. Create a column calculating the percentage of minority students by subtracting the number of white students enrolled from the total amount of students.
11. Create csv from the data frame.

The data for median income was downloaded as a csv from the United States Census Bureau site. The federal funding for each school district was downloaded as a csv from the New York State 2019 District data as FederallyFundedDistricts.csv.

The zip code to district mapping data is from the NCES geographic files dataset for 2021. Specifically, the file matches Local Education Area codes (LEAs) with Zip Code Tabulation Areas (ZCTAs). The zip code to NYC school district mapping data is created manually by combining the NYC Zip Code Tabulation Areas (NYC OpenData, 2020) GIS file with the NYC School Districts (NYC OpenData, 2022) GIS File in mapshaper.org and manually recording all overlaps between NYC zip codes and school districts.

The map data comes from the NYS GIS Dataset, under NYS Schools and School District Boundaries (NYS GIS, 2019). This dataset is not transformed in Python because it is a GIS file. It is transformed in mapshaper.org, a free browser GIS tool.

Map (in Power BI) (check the map folder)

The map visualization in Power BI is created using the custom Shape Map visualizations, which is currently a preview feature. In case the Shape Map visualization does not show up for you in the tab, go to **File > Options and Settings > Options > Preview Features** and select the **Shape**

map visual checkbox. The documentation for Shape Map visualizations is available on the Microsoft website (Microsoft, 2020).

The Shape Map visualization with a custom map requires GIS data in the form of the TopoJSON format. The New York State School District Boundaries Shapefile (NYC Education Department, 2019) was converted into a custom map through the following process. The documentation page suggests using tools like https://mapshaper.org/ to convert GIS data of other types into TopoJSON format. Additionally, Power BI will not display the TopoJSON file correctly unless converted into WGS 84, a GIS standard that Power BI accepts. To run this conversion in mapshaper, open the console and run the command **-proj wgs84**. This solution is referenced from a Power BI community thread (PowerBI Community Forum, 2021). The final map TopoJSON file is imported as a custom map using the Equirectangular projection.

To match our custom map to the data, Power BI takes the features in the TopoJSON file of each separate region and uses them as keys to match with the location section of the Shape Map's data input. You can check the features of each region in by mousing over each region in https://mapshaper.org/. The feature that best matches our cleaned data is the "POPULAR_NA", or the Popular Name of the school districts. In the "Transform" data cleaning steps, the names of each school district in our dataset have been changed to reflect the NYS Education Department's naming conventions. When some districts in the NYS GIS data do not match their own naming conventions, they were changed directly using the "edit attributes" feature in https://mapshaper.org/. This means that Power BI can match our dataset to the map given the "DistrictName" column of our dataset is applied to the "Location" section of the map input.

Transformation

Did you use all of the data you extracted as-is? Did you remove columns? Did you change columns' names? Did you change your column formats? What steps were taken to get the data in a form that you could use it? Be sure to number steps when the order matters.

1. Median Income to School District (Census): Median Income to Zip Code to School District
    a. Median income data sourced from the US Census Bureau was cleaned for just the Geographic Area Name and the Estimated Household Median income (dollars).
        i. The Geographic Area Name was converted to just zip codes as integers.
        ii. The column for Geographic Area Name is renamed to be ZipCode, and Estimated Household Median income (dollars) is renamed to be MedianIncome for later merging.
        iii. The MedianIncome column is coerced into numerical data types. This results in an error because some of the data is in string format due to median income exceeding a certain number, such as "250,000+".

        iv. This table is then dumped into the SQL database before being later merged.
    b. Zip code to school district data sourced from the NCES Geographical Relationship files was filtered for just the columns for name of the school district (NAME_LEA21) and the zip code (ZCTA5CE20).
        i. Because the dataset is for the entire United States, the zip codes (ZCTA5CE20) were filtered for just those in New York State (6390 and all numbers between 10001 and 14905).
        ii. Since all NYC school districts in this dataset are bundled together into the "New York City Department of Education," and we are looking for each separate geographic district in the NYC school system, the combined "New York City Department of Education" is filtered out and removed
        iii. To comply with naming conventions with the rest of the data, the school district names (NAME_LEA21) are all converted to uppercase.
        iv. For merging into the median income dataset later on, the columns are renamed from NAME_LEA21 to DistrictName, and from ZCTA5CE20 to ZipCode.
    c. The zip code to school district dataset for just NYC school districts consists of the school district number (District) and the zip code (ZipCode).
        i. To comply with naming conventions with the rest of the data, a district name map dictionary for each school district number to district name is created and applied to the District column.
        ii. To merge onto the other zip code file, the District column is renamed to DistrictName.
        iii. This table is then dumped into the SQL database before being later merged.
    d. The NYS and NYC zip code to school district datasets are concatenated using pd.concat().
    e. This table is then dumped into the SQL database before being later merged.
    f. The combined NYS and NYC zip code to school district mapping data frame are merged onto the median income dataset using the ZipCode columns as a key as a Left join. This is so school districts with no data (likely due to the coerce errors when cleaning median income, or because there are missing values in the census data) will still be listed in the dataset for later processing.
    g. The median income is grouped by DistrictName for each school district, and the average median income is found across all the zip codes that fall into each school district. This resulting dataset of school districts to zip codes is exported to csv.
2. Web Scraping and Data Cleaning
    a. The data from the NYSED web scraped data was cleaned by removing irrelevant symbols and punctuation from each of the value lists. The district names were cleaned by using rstrip. All rows that contain "N/A" in the WhiteStudentsEnrolled

column were removed. 5 districts that reported a graduation rate of 0 were removed due to them skewing the data.

   b. The median income data was cleaned by renaming the column headers to be clear. The values in the zip code column were stripped to include only the zip code value.
   c. In the NYSED code the median income data had some of the district names reformatted to match the original district names.
      i. "CENTRAL SCHOOL DISTRICT" to "CSD"
      ii. "UNION FREE SCHOOL DISTRICT" to "UFSD"
   d. The median income data frame was then merged with the NYSED data on district name.
   e. Save data frame as "PreliminaryDistrictDataWITHMEDIANINCOME2.csv"
   f. This table is then dumped into the SQL database before being later merged.
3. NYC District Data:
   a. The FundingPerStudent column is found for each NYC school district by dividing the Total Budget Allocation (from the NYC Districts Funding dataset) by the Total Enrollment per district.
   b. Save dataframe as "PreliminaryDistrictDataNYCDistrictsFunding3.csv"
4. Federal Funding and Naming Conventions
   a. Remove unnecessary columns from FederallyFundedDistricts.csv
   b. Change columns to match the district names.
   c. Load in PreliminaryDistrictDataNYCDistrictsFunding3.csv and update its district names to match PowerBI map Json naming conventions. Update district names for merging.
   d. Merge the Federal funding data frame and the district data frame.
   e. Change values of federal funding to "Yes" for having federal funding and "No" if there is no federal funding. Update all NYC districts to be "Yes"
   f. Convert Graduation rate and Dropout rate to be float and change "-" to "NaN"

The school district zip code data set holds the zip code to school district. The district data set holds variables that involve the demographic variables of each school district, the district names, graduation rates, dropout rates, funding per student, student enrollment numbers, and minority percentage.

Load

Each dataset is given a meaningful name-- school_district_zip_code, DistrictDataCleaned, and PreliminaryDistrictDataWITHMEDIANINCOME--and saved as CSVs in the data factory. The data was then sent to the Kafka server to simulate streaming data. All datasets, the Kafka producer and consumer code, and the ETL Jupyter Notebooks can be found in the GitHub repository.

Conclusion

Following ETL process create visualizations. Exploratory data analysis and visualizations are to be completed with Python using Pandas, Matplotlib, and Seaborne.

References

Create Shape Map Visualizations in Power BI Desktop (preview). Use Shape maps in Power BI Desktop (Preview) - Power BI | Microsoft Docs. (2022). Retrieved from https://docs.microsoft.com/en-us/power-bi/visuals/desktop-shape-map

Department of Health and Mental Hygiene (DOHMH). (2020, May 13). *Modified ZIP code tabulation areas (MODZCTA)*. Modified Zip Code Tabulation Areas (MODZCTA) | NYC Open Data. From https://data.cityofnewyork.us/Health/Modified-Zip-Code-Tabulation-Areas-MODZCTA-/pri4-ifjk

NCES. (2022). *School District Geographic Relationship Files*. Geographic. Retrieved from https://nces.ed.gov/programs/edge/Geographic/RelationshipFiles "\GRF21\grf21_lea_zcta5ce20.xlsx"

New York State. (2019). *2019 District Data – School Funding Transparency*. https://openbudget.ny.gov/sft/sft-districts-19.html

New York State. (2022). *GIS.NY.GOV*. NYS GIS Clearinghouse - NYS Education Department - NYS Schools and School District Boundaries. From https://gis.ny.gov/gisdata/inventories/details.cfm?DSID=1326

NYC Open Data. (2022). *School districts*. NYC Open Data. From https://data.cityofnewyork.us/Education/School-Districts/r8nu-ymqj

NYC OpenData. (2021). *FY2020 Local Law 16 Final Report*. https://data.cityofnewyork.us/Education/FY2020-Local-Law-16-Final-Report/cvqn-xqrr/data

NYSED. (2020). *Districts: NYSED Data Site*. data.nysed.gov. From https://data.nysed.gov/lists.php?type=district

Problem Using Custom Shape Map. Microsoft Power BI Community. (2021, September 13). Retrieved from https://community.powerbi.com/t5/Desktop/Problem-using-custom-shape-map/m-p/55795

U.S. Census Bureau. (2020). *Income in the Past 12 Months (In 2020 Inflation-Adjusted Dollars)*. https://data.census.gov/cedsci/table?q=median+income&g=860XX00US11701